DOCUMENT RESUME

ED 408 302 TM 026 504

AUTHOR Snyder, Patricia A.; Thompson, Bruce
TITLE Use of Tests of Statistical Significance and Other Analytic Choices in a School Psychology Journal: Review of Practices and Suggested Alternatives.
PUB DATE 24 Jan 97
NOTE 25p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, January 24, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Educational Psychology; *Educational Research; *Effect Size; Elementary Secondary Education; Research Methodology; Research Reports; *Scholarly Journals; *School Psychologists; Statistical Inference; *Statistical Significance; Test Interpretation; *Test Use
IDENTIFIERS American Psychological Association

ABSTRACT

The use of tests of statistical significance was explored, first by reviewing some criticisms of contemporary practice in the use of statistical tests as reflected in a series of articles in the "American Psychologist" and in the appointment of a "Task Force on Statistical Inference" by the American Psychological Association (APA) to consider recommendations leading to improved practice. Related practices were reviewed in seven volumes of the "School Psychology Quarterly," an APA journal. This review found that some contemporary authors continue to use and interpret statistical significance tests inappropriately. The 35 articles reviewed reported a total of 321 statistical tests for which sufficient information was provided for effect sizes to be computed, but authors of only 19 articles did report various magnitudes of effect indices. Suggestions for improved practice are explored, beginning with the need to interpret statistical significance tests correctly, using more accurate language, and the need to report and interpret magnitude of effect indices. Editorial policies must continue to evolve to require authors to meet these expectations. (Contains 50 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

schlpsyc.wp1 3/7/97

ED 408 302

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
BRUCE THOMPSON
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Use of Tests of Statistical Significance and Other Analytic Choices in a School Psychology Journal: Review of Practices and Suggested Alternatives

Patricia A. Snyder
LSU Medical Center 70019-2799

Bruce Thompson
Texas A&M University 77843-4225
and
Baylor College of Medicine

Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX, January 24, 1997.

TM026504

Use of Tests of Statistical Significance and Other Analytic Choices in a School Psychology Journal: Review of Practices and Suggested Alternatives

ABSTRACT

The present work had three purposes. First, some of the criticisms of contemporary practice as regards the use of statistical tests are briefly reviewed; these concerns have been reflected in a series of articles in the American Psychologist and in the appointment of an American Psychological Association (APA) Task Force on Statistical Inference which will consider recommendations leading to improved practice as regards the use of statistical significance tests. Second, related practices within seven volumes of an APA journal, School Psychology Quarterly, are reviewed; it was found that some contemporary authors continue to use and interpret statistical significance tests inappropriately. Third, suggestions for improved practice are briefly explored.

The Board of Scientific Affairs within the American Psychological Association (APA), following nearly two years of discussion, has now appointed an APA Task Force on Statistical Inference. The Task Force has a distinguished membership (e.g., Robert Rosenthal, Co-Chair, and Jacob Cohen, Co-Chair), as well as a distinguished advisory panel (i.e., Lee Cronbach, Paul Meehl, Fred Mosteller, and John Tukey).[1] As described in some detail by Shea (1996), the Task Force is studying current uses of statistical significance tests within APA journals and other outlets.

The Task Force was created following the recent publication of a series of articles in the American Psychologist (Cohen, 1990; Kupfersmid, 1988; Rosenthal, 1991; Rosnow & Rosenthal, 1989); particularly influential have been recent articles by Cohen (1994), Kirk (1996), Schmidt (1996), and Thompson (1996). The entire Volume 61, Number 4 issue of the Journal of Experimental Education was devoted to these themes.[2]

These recent works followed a numerous previous calls for improved research practice that have been published throughout the last 35 years. Particularly noteworthy among these have been the publications by Rozeboom (1960), Morrison and Henkel (1970), Meehl (1978), Shaver (1985), and especially Carver (1978).

The present work has three purposes. First, some of the criticisms of contemporary practice as regards the use of statistical tests are briefly reviewed. Second, related practices within seven volumes of an APA journal, School Psychology Quarterly, are reviewed. Third, suggestions for improved practice

are briefly explored.

Three Criticisms of Contemporary Practice

Three among the various possible criticisms of the ways that many researchers use statistical significance tests as interpretation aids will be noted here. As is often the case, some of these problems involve the ways that researchers use their tools, rather than inherent problems with the tools themselves.

Use of p as an Evaluation of Result Replicability

Many researchers vest statistical tests with exaggerated importance because they incorrectly believe that p values evaluate the probability that sample results occur (or the null hypothesis is false) in the population. Such a result would be noteworthy, if that was what statistical significance tested, but these tests simply do not test for population values.

A test of the population would be noteworthy, because if we knew more about population then we would know more about what other researchers might find in future samples drawn from the population. The classic example of belief in the fallacy that statistical significance tests the population is provided by Melton (1962), who after 12 years as editor of the Journal of Experimental Psychology stated that:

> In editing the Journal there has been a strong reluctance to accept and publish results related to the principal concern of the researcher when those results were [statistically] significant [only] at the .05 level... It reflects a belief that it is the

> responsibility of the investigator in a science to reveal his [sic] effect in such a way that no reasonable man [sic] would be in a position to discredit the results by saying that they were the product of the way the ball bounces. (p. 554)

Statistical significance tests do <u>not</u> compute the probability of population results, given the sample results. Instead, as various authors (see especially Cohen (1994) and Thompson (1996)) have emphasized, statistical significance tests evaluate the probability of the *sample* values, *assuming* that the null hypothesis is exactly descriptive of the *population*. This second issue is somewhat less interesting.

The two statements are not the same. The two elements in the logic (population values and sample values) are the same, but which values are taken as givens are inherently different, and this difference means that the two interpretations are irreconcilable.

Put simply, the direction of statistical inference in statistical significance tests is *from* the population *to* the sample, and <u>not</u> from the sample to the population. As eloquently explained by Cohen (1994), the test of the conventional null hypothesis

> ...does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does! What we want to know is "Given these [sample] data, what is the probability that H_0 is true [in the

> population]?" But as most of us know, what it tells us is "Given that H_0 is true [in the population], what is the probability of these (or more extreme) [sample] data?" These are not the same.... (p. 997)

This discussion should not be taken as implying that result replicability is unimportant. To the contrary, science proceeds by cumulating evidence that particular results occur under stated conditions. What is said here is that statistical significance tests do not (do not, do not...) evaluate result replicability. Other analyses, such as so-called "external" or "internal" replicability analyses (e.g., cross-validation, jackknife, bootstrap), must and should be used as interpretation aids for this purpose (cf. Thompson, 1994b, 1995a, 1996).

Use of p as a Measure of Result Importance

One problematic aspect of statistical significance tests is that researchers almost always use null hypotheses of no difference or of zero relationship. When such hypotheses are used, and zero population effects are thereby assumed to be exactly descriptive of the population, p values are calculated on the basis of a premise that we know to be false (see Thompson, 1996). And a false premise renders at least somewhat inaccurate any conclusions deduced from that premise.

Various prominent statisticians have long acknowledged that the null hypothesis of no difference is never true in the population (Tukey, 1991). Consequently, there will always be some differences in population parameters, although the differences may

be incredibly trivial. Some 40 years ago Savage (1957, pp. 332-333) noted that, "Null hypotheses of no difference are usually known to be false before the data are collected." Subsequently, Meehl (1978, p. 822) argued, "As I believe is generally recognized by statisticians today and by thoughtful social scientists, the null hypothesis, taken literally, is always false." Similarly, statistician Hays (1981, p. 293) pointed out that "[t]here is surely nothing on earth that is completely independent of anything else. The strength of association may approach zero, but it should seldom or never be exactly zero."

This realization means that non-zero sample effects are always expected, and that consequently "virtually any study can be made to show [statistically] significant results if one uses enough subjects" (Hays, 1981, p. 293). As Nunnally (1960, p. 643) noted, "If the null hypothesis is not rejected, it is usually because the N is too small. If enough data are gathered, the hypothesis will generally be rejected."

It is important to understand that because the null hypothesis of no difference is always false, every study will achieve statistical significance at some sample size. This realization means that statistical significance tests are neither tests of result replicability nor pure measures of result importance; the tests largely measure researcher endurance. As Thompson (1992) noted:

> Statistical significance testing can involve a tautological logic in which tired researchers,

> having collected data from hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and know they're tired. This tautology has created considerable damage as regards the cumulation of knowledge... (p. 436)

Use of Better Language

Thompson (1996) recommended that when the null hypothesis is rejected, "such results ought to always be described as 'statistically significant,' and should never be described only as 'significant'" (pp. 28-29). The argument was that the common meaning of "significant" as "important" has nothing to do with the statistical use of this term, because statistical significance does not measure importance (a) in the form of replicability or (b) in the form of noteworthiness (see Carver, 1993; Shaver, 1985).

Several methodologists have argued that the use of the complete phrase, "statistically significant" as against "significant", might help to convey to at least some readers of research that the use of this technical term has a different meaning not connoting result importance. Carver (1993) eloquently made the argument:

> When trying to emulate the best principles of science, it seems important to say what we mean and to mean what we say. Even though many readers of scientific journals know that the word *significant*

> is supposed to mean *statistically significant* when it is used in this context, many readers do *not* know this. Why be unnecessarily confusing when clarity should be most important? (p. 288, emphasis in original)

The fact that more thoughtful or more highly trained readers will know the correct meaning of the telegraphic wording does not excuse gratuitously confusing lay readers or student readers who are only beginning their training.

This discussion does not mean that result importance should be ignored, but is meant to emphasize that improbable sample results assuming a false premise are not necessarily important. Importance of results can be evaluated, but magnitude of effect indices must be evoked for this purpose. Snyder and Lawson (1993) reviewed several of the many alternatives for evaluating result importance.

Contemporary Practices in a School Psychology Journal

The School Psychology Quarterly is published as the official journal of APA Division 16 (School Psychology). The journal began in 1986 under the title, Professional School Psychology. The name change was implemented in 1989 to convey a broader focus to include more research reports. The journal has had two editors over the course of its first 10 year history (1986-1996).

The present review examined the use of statistical significance tests and other analytic choices within the 35 research articles published in School Psychology Quarterly volumes 5 through 11. For each research article published in the volumes

examined, we recorded (a) the research topic and data collection method; (b) the sample size; (c) the statistical analyses used; (d) whether and how statistical significance was reported; (e) whether and how a magnitude of effect index (Snyder & Lawson, 1993) was reported; (f) whether and how an "external" or an "internal" replicability analysis was conducted, and (g) whether other interpretation aids such as confidence intervals or standard errors were used. Exceptional features of analytic practice were also noted.

The 35 articles reported a total of 321 statistical tests for which sufficient information was provided for effect sizes to be computed (in various cases authors did not report sufficient information to compute effect sizes for results that were not statistically significant). The mean of these 321 effect sizes was .13 (SD = .16); this value is comparable to the effect that Cohen (1988) characterized as "medium" or average across various literatures. A total of 192 of these tests were statistically significant. Several conclusions can be extrapolated from our results.

First, regarding language use, authors of only five of the 35 articles used the term "statistically significant" rather than "significant" (Fuchs, Fuchs, Harris & Roberts, 1996; Hyatt & Tingstrom, 1993; Kieth & Cool, 1992; MacMann & Barnett, 1994; Turner, Biedel, Hughes & Turner, 1993). This pattern is somewhat troublesome, for the reasons cited earlier. On the other hand, no authors referred to results as being "highly significant." Only

one of the articles made other classic mistakes in language. In that article the authors referred to results as "approaching significance"; Thompson (1993) commented thusly on this language use:

> ...one fellow editor I know will not tolerate sloppy speaking regarding statistical tests. Whenever authors note in a manuscript that "the results *approached* statistical significance", he always immediately writes the authors back with the query, "How do you know your results were not working very hard to *avoid* being statistically significant?" (p. 285)

Second, authors of 19 articles did report various magnitude of effect indices (e.g., Kratochwill, Elliot & Busse, 1995). However, even among these authors, few authors interpreted these indices. For example, in several articles squared correlation coefficients were reported but not interpreted. On the other hand, some authors noted that their statistically significant results should be interpreted with caution given the value of eta^2 (i.e., one magnitude of effect index).

The preponderance of the authors emphasized tests of statistical significance to determine if their results were noteworthy. What was particularly dramatic was that some of these studies were overinterpreted (i.e., studies with small effects but large sample sizes--Norris, Burke & Speer, 1990) while other results were underinterpreted (i.e., studies with large effects but

small sample sizes--Fuchs et al., 1996). Such are the vagaries resulting from misinterpretation of statistical significance tests.

Third, authors of only 2 of the 35 articles invoked an "internal" replicability analysis, such as cross-validation, the jackknife, or the bootstrap (Elias & Allen, 1991); Kieth & Cool, 1992). In only two studies did authors conduct an actual "external" replication with an independent sample of new subjects (Jorgenson, Jorgenson, Gillis & McCall, 1993; Vickers & Minke, 1995). Again, authors who think that statistical significance evaluate result replicability will erroneously find such replicability analyses less necessary, with all the attendant negative consequences for the business of accurately cumulating evidence across studies.

Fourth, almost all authors who failed to reject their null hypotheses did not conduct power analyses to determine whether their results were artifacts of small sample size. An exception was the study reported by Hughes, Grossman and Barker (1992), who described at what sample size their non-statistically significant results would have been statistically significant. Persons who vest confidence in the statistical significance test logic should be expected to conduct power analyses when results for important hypotheses are not statistically significant.

Three other patterns incidental to the primary focus of our work also must be noted. First, many of the authors who used regression methods elected to use stepwise methods (e.g., Huebner, 1991, 1992; Jorgenson et al., 1993). The pattern is regrettable,

because methodologists are critical of stepwise methods, since these methods yield distorted and non-replicable findings (see Huberty (1989), Snyder (1991), and especially Thompson (1995b)).

As Cliff (1987, p. 185) noted, "most computer programs for [stepwise] multiple regression are positively satanic in their temptations toward Type I errors." He also suggested that, "a large proportion of the published results using this method probably present conclusions that are not supported by the data" (pp. 120-121).

Second, several authors used series of univariate tests to evaluate separately each dependent variable in large sets of dependent variables (e.g., Cowen, Pryor-Brown, Hightower & Lotyczewski, 1991; Norris, Burke & Speer, 1990). This practice leads to inflation of experimentwise error rates and also may distort the reality about which the researcher is attempting to generalize (Thompson, 1992, 1994c).

Even among authors who used multivariate analyses, many of these authors used univariate tests as post hoc methods to understand their multivariate effects. This practice is incorrect. As noted elsewhere:

> The "protected F-test" analytic approach is inappropriate and wrong-headed. ...*[U]nivariate post hoc tests do not inform the researcher about the differences in the multivariate latent variables actually analyzed in the multivariate analysis*. It is illogical to first declare interest in a

> multivariate omnibus system of variables, and to then explore detected effects in this multivariate world by conducting non-multivariate tests! (Thompson, 1994c, p. 14, emphasis in original)

Third, authors of none of the articles followed recommendations by Carver (1993) and others to report interpretation aids such as estimates of sampling error (e.g., confidence intervals and standard errors). Use of such aids might help remind readers that tests of statistical significance are fallible point estimates.

Suggestions for Improvement

Some 45 years ago, prominent statistician Yates (1951, pp. 32-33) suggested that the use of statistical significance tests

> ...has caused scientific research workers to pay undue attention to the results of the tests of [statistical] significance they perform on their data, and too little to the estimates of the magnitude of the effects they are investigating... The emphasis on tests of [statistical] significance, and the consideration of the results of each experiment in isolation, have had the unfortunate consequence that scientific workers have often regarded the execution of a test of [statistical] significance on an experiment as the ultimate objective.

And Meehl (1978, p. 817, 823) argued some 15 years ago:

> I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft [i.e., social science] areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology... I am not making some nit-picking statistician's correction. I am saying that the whole business is so radically defective as to be scientifically almost pointless.

Two things are needed to overcome the inertia reflected in decades of refusals (a) to correctly interpret statistical significance tests when they are used, (b) to use better language regarding these tests, and (c) to always report and interpret magnitude of effect indices (e.g., eta^2, $omega^2$, R^2), and (d) to always evaluate result replicability in some way. First, more researchers must confront a hesitancy to understand genuinely what statistical tests do and do not do.

Second, editorial policies must continue to evolve to require authors to meet the expectations presented here. Some incremental progress was made when the fourth edition of the APA style manual was revised to note that:

> Neither of the two types of probability values reflects the importance or magnitude of an effect because both depend on sample size... You are encouraged to provide effect-size information. (APA,

1994, p. 18)

Of course, it has been argued that the reporting and interpretation of effect sizes should have been *required* rather than merely encouraged (Thompson, 1996).

Certainly, some journal editorial boards have revised editorial policies to reflect contemporary thinking as regards statistical significance tests. For example, the guidelines for authors of Measurement and Evaluation in Counseling and Development have for several years noted that:

> 7. Authors are strongly encouraged to provide readers with effect size estimates as well as statistical significance tests.... 8. Studies in which statistical significance is not achieved will still be seriously considered for publication.... (Association for Assessment in Counseling, 1990, p. 48)

Similarly, the author guidelines for Educational and Psychological Measurement *require* authors to report and interpret effect sizes, and strongly encourage authors to report actual "external" replication studies, or to conduct "internal" replicability analyses. Regarding language use, these guidelines also provide that, "We will follow the admonitions of others... [by proscribing] the use of only the words, 'significant' or 'significance', when referring to statistical significance" (Thompson, 1994a, p. 844).

The revised author guidelines of the Journal of Experimental

Education also address some of these issues. The new guidelines for contributors state:

> In consideration of contemporary thinking about statistical significance tests, reflected in the 1993 JExE theme issue (Vol. 61, No. 4), authors are encouraged to use the phrase "statistical significance" rather than only "significance" whenever referring to the results of inferential tests. Furthermore, authors are required to report and interpret magnitude-of-effect measures in conjunction with every p value that is reported... (Heldref Foundation, in press)

Hopefully, as researchers and board members reflect on their practices, more and more editorial boards will formulate more informed policies as regards the issues presented here. The clients we serve from within our professions deserve best practice as regards reporting and interpreting research that informs our intervention decisions.

Footnotes

[1]The core members of the APA Task Force on Statistical Inference are: Bob Rosenthal, Chair, Robert Abelson, and Jacob Cohen. Other members of the Task Force are: Leona Aiken, Mark Applebaum, Gwen Boodoo, David Kenny, Helena Kramer, Don Rubin, Bruce Thompson, Howard Wainer, and Lee Wilkinson. Professors Lee Cronbach, Paul Meehl, Fred Mosteller, and John Tukey are serving as advisors to the Task Force.

[2]Interested readers may request a gratis copy of this theme issue by e-mailing a request (including a postal address) to Professor Thompson at E100BT@TAMVM1.TAMU.EDU.

References

American Psychological Association. (1994). Publication manual of the American Psychological Association (4th ed.). Washington, DC: Author.

Association for Assessment in Counseling. (1990). Guidelines for authors. Measurement and Evaluation in Counseling and Development, 23(1), 48.

Carver, R. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Carver, R. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61, 287-292.

Cliff, N. (1987). Analyzing multivariate data. San Diego: Harcourt Brace Jovanovich.

Cohen, J. (1988). Statistical power analysis (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45(12), 1304-1312.

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 49, 997-1003.

Cowen, E. L., Pryor-Brown, L., Hightower, A. D., & Lotyczewski, B. S. (1991). Age perspectives on the stressfulness of life-events for 10-12 year old children. School Psychology Quarterly, 6(4), 241-250.

Elias, M. J., & Allen, G. J. (1991). A comparison of instructional methods for delivering a preventive social competence/social

decision making program to at risk, average, and competent students. School Psychology Quarterly, 6(4), 251-272.

Fuchs, D., Fuchs, L. S., Harris, A. H., & Roberts, P. H. (1996). Bridging the research-to-practice gap with mainstream assistance teams: A cautionary tale. School Psychology Quarterly, 11(3), 244-266.

Hays, W. L. (1981). Statistics (3rd ed.). New York: Holt, Rinehart and Winston.

Heldref Foundation. (in press). Guidelines for contributors. Journal of Experimental Education.

Huberty, C. (1989). Problems with stepwise methods--better alternatives. In B. Thompson (Ed.), Advances in social science methodology (Vol. 1, pp. 43-70). Greenwich, CT: JAI Press.

Huebner, E. S. (1991). Correlates of life satisfaction in children. School Psychology Quarterly, 6(2), 103-111.

Huebner, E. S. (1992). Burnout among school psychologists: An exploratory investigation into its nature, extent, and correlates. School Psychology Quarterly, 7(2), 129-136.

Hughes, J. N., Grossman, P., & Barker, D. (1990). Teachers' expectancies, participation in consultation, and perceptions of consultant helpfulness. School Psychology Quarterly, 5(3), 167-179.

Hyatt, S. P., & Tingstrom, D. H. (1993). Consultants' use of jargon during intervention presentation: An evaluation of presentation modality and type of intervention. School Psychology Quarterly, 8(2), 99-109.

Jorgenson, C. B., Jorgenson, D. E., Gillis, M. K., & McCall, C. M. (1993). Validation of a screening instrument for young children with teacher assessment of school performance. School Psychology Quarterly, 8(2), 125-139.

Keith, T. Z., & Cool, V. A. (1992). Testing models of school learning: Effects of quality of instruction, motivation, academic coursework, and homework on academic achievement. School Psychology Quarterly, 7(3), 207-226.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56(5), 746-759.

Kratochwill, T. R., Elliott, S. N., & Busse, R. T. (1995). Behavior consultation: A five-year evaluation of consultant and client outcomes. School Psychology Quarterly, 10(2), 87-117.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. American Psychologist, 43, 635-642.

MacMann, G. M., & Barnett, D. W. (1994). Structural analysis of correlated factors: Lessons from the verbal-performance dichotomy of the Wechsler Scales. School Psychology Quarterly, 9(3), 161-197.

Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Melton, A. (1962). Editorial. Journal of Experimental Psychology, 64, 553-557.

Morrison, D.E., & Henkel, R.E. (Eds.). (1970). The significance

test controversy. Chicago: Aldine.

Norris, D. A., Burke, J. P., & Speer, A. L. (1990). Tri-level service delivery: An alternative consultation model. School Psychology Quarterly, 5(2), 89-110.

Nunnally, J. (1960). The place of statistics in psychology. Educational and Psychological Measurement, 20, 641-650.

Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the BESD, and alternative indices. American Psychologist, 46, 1086-1087.

Rosnow, R.L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276-1284.

Rozeboom, W.W. (1960). The fallacy of the null hypothesis significance test. Psychological Bulletin, 57, 416-428.

Savage, R.J. (1957). Nonparametric significance. Journal of the American Statistical Association, 52, 331-344.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. Psychological Methods, 1(2), 115-129.

Shaver, J. (1985). Chance and nonsense. Phi Delta Kappan, 67(1), 57-60.

Shea, C. (1996). Psychologists debate accuracy of "significance test." Chronicle of Higher Education, 42(49), A12, A16.

Snyder, P. (1991). Three reasons why stepwise regression methods should not be used by researchers. In B. Thompson (Ed.), Advances in educational research: Substantive findings,

methodological developments (Vol. 1, pp. 99-105). Greenwich, CT: JAI Press.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. Journal of Experimental Education, 61(4), 334-349.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling and Development, 70, 434-438.

Thompson, B. (1993). Foreword. Journal of Experimental Education, 61, 285-286.

Thompson, B. (1994a). Guidelines for authors. Educational and Psychological Measurement, 54(4), 837-847.

Thompson, B. (1994b). The pivotal role of replication in psychological research: Empirically evaluating the replicability of sample results. Journal of Personality, 62(2), 157-176.

Thompson, B. (1994c, February). Why multivariate methods are usually vital in research: Some basic concepts. Paper presented as a Featured Speaker at the biennial meeting of the Southwestern Society for Research in Human Development, Austin, TX. (ERIC Document Reproduction Service No. ED 367 687)

Thompson, B. (1995a). Exploring the replicability of a study's results: Bootstrap statistics for the multivariate case. Educational and Psychological Measurement, 55, 84-94.

Thompson, B. (1995b). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial.

Educational and Psychological Measurement, 55, 525-534.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25(2), 26-30.

Tukey, J.W. (1991). The philosophy of multiple comparisons. Statistical Science, 6, 100-116.

Turner, B. G., Beidel, D. C., Hughes, S., & Turner, M. W. (1993). Test anxiety in African American school children. School Psychology Quarterly, 8(2), 140-152.

Vickers, H. S., & Minke, K. M. (1995). Exploring parent-teacher relationships: Joining and communication to others. School Psychology Quarterly, 10(2), 133-150.

Yates, F. (1951). The influence of Statistical methods for research workers on the development of the science of statistics. Journal of the American Statistical Association, 46, 19-34.